Annotating Data Sets

27 December 2013
NJG

Annotation of data sets with proper meta-data is essential. The best way to do this is to embed the meta-data in the data file itself. In R, this is easy to do by using the `#` comment and then whatever follows in that line will be skipped when R reads in the data. Here is a minimal set of meta-data

```
# Author: Nicholas J. Gotelli
# Date: 27 December 2013
#
# Data Set: A short one-sentence descriptor of what the data are
# Data Source: Publication or other listing for where these data came from or are stored
# Funding Source: Any grant or other attributes that are needed
#
# Data Collection: Details on collection of the data (may take multiple lines)
# Rows: Details on the row elements (may take multiple lines)
# Columns: Details on the columns, including measurement units (may take multiple lines)
Species, C1, C2, C3
Species1, 1, 1, 0
Species2, 1, 0, 0
...
```

R should be able to read this, but if you try the following, the file will not be read properly:

```
read.csv("Input_File.csv",header=TRUE,row.names=1)
```

Perversely, the `read.csv` command has the comment character disabled. You could do this by skipping the proper number of lines:

```
read.csv("Input_Data.csv",header=TRUE,row.names=1,skip=10)
```

This will indeed skip the first 10 lines of this file and begin reading where it should. However, each data set would need to be hand-wired in this way, which is problematic for batch processing.

A better solution is to use the `read.table` command, and insert the needed delimiter:

```
read.table("Input_File.csv",header=TRUE,row.names=1,sep=",")
```

In this way, all of the comments are skipped, and you don't have to worry about how many lines of meta-data are contained in the file.